

HRTF-BASED TWO-DIMENSIONAL ROBUST LEAST-SQUARES FREQUENCY-INVARIANT BEAMFORMER DESIGN FOR ROBOT AUDITION

Hendrik Barfuss, Michael Buerger, Jasper Podschus, and Walter Kellermann

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg
Cauerstr. 7, 91058 Erlangen, Germany

{hendrik.barfuss, michael.buerger, jasper.podschus, walter.kellermann}@fau.de

ABSTRACT

In this work, we propose a two-dimensional Head-Related Transfer Function (HRTF)-based robust beamformer design for robot audition, which allows for explicit control of the beamformer response for the entire three-dimensional sound field surrounding a humanoid robot. We evaluate the proposed method by means of both signal-independent and signal-dependent measures in a robot audition scenario. Our results confirm the effectiveness of the proposed two-dimensional HRTF-based beamformer design, compared to our previously published one-dimensional HRTF-based beamformer design, which was carried out for a fixed elevation angle only.

Index Terms— Spatial filtering, robust superdirective beamforming, white noise gain, signal enhancement, robot audition

1. INTRODUCTION

Spatial filtering approaches are an effective means to acoustically focus on a target source whose emitted sound waves impinge from a certain Direction of Arrival (DoA). When the microphone array is mounted on a humanoid robot's head, spatial filtering algorithms should take the effect of the robot's head on the sound field into account, so that an adequate signal enhancement performance can be expected [1]. One possibility to achieve this is to incorporate the Head-Related Transfer Functions (HRTFs)¹ of the robot's head as steering vectors into the beamformer design, see, e.g., [2, 3].

Following this strategy, we presented a data-independent HRTF-based Robust Least-Squares Frequency-Invariant (RLSFI) beamformer design in [1], which is based on the work by Mabande et al. [4, 5]. In addition to using HRTFs as steering vectors, the beamformer design allows the user to directly control the White Noise Gain (WNG) and, therefore, the beamformer's robustness against microphone self-noise, microphone mismatch or mis-positioning of microphones, see, e.g., [6, 7, 8].

However, one major drawback of the beamformer design in [1] is its limitation to a plane corresponding to a fixed elevation angle, which turned out to be inappropriate for capturing a three-dimensional sound field, especially when considering that a robot head changes its elevation angle relative to the target. Therefore,

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

¹Please note that even though the robot's head does not have pinnae and the microphone positions are not limited to 'ear positions', we still use the term HRTF here. Furthermore, in the context of this work, HRTFs only model the direct propagation path between a source and a microphone mounted on the robot's head, but no reverberation components.

in this work, we extend the beamformer design of [1] such that the beamformer response can be controlled for all DoAs on a sphere surrounding the humanoid robot. Note that similar problem has been investigated in [?] for a linear four-element microphone array employed in the Microsoft Kinect™ using a Minimum-Variance Distortionless Response (MVDR) beamformer.

The remainder of this article is structured as follows: In Section 2 the HRTF-based RLSFI beamformer design is introduced, and the proposed extension to two dimensions is motivated and presented. An evaluation of the extended HRTF-based beamformer design is presented in Section 3. Finally, conclusions and an outlook to future work are given in Section 4.

2. HRTF-BASED ROBUST BEAMFORMING FOR ROBOT AUDITION

2.1. HRTF-based robust least-squares frequency-invariant beamformer design

The block diagram of a time-domain Filter-and-Sum Beamformer (FSB) with N channels is illustrated in Fig. 1. The output signal $y[k]$ at time instant k is obtained by a convolution of the microphone signals $x_n[k]$, $n \in \{0, \dots, N-1\}$ with Finite Impulse Response (FIR) filters $\mathbf{w}_n = [w_{n,0}, \dots, w_{n,L-1}]^T$ of length L , followed by a summation over all channels. The beamformer response of an FSB as depicted in Fig. 1 is given by [4, 9]:

$$B(\omega, \phi, \theta) = \sum_{n=0}^{N-1} W_n(\omega) g_n(\omega, \phi, \theta), \quad (1)$$

where $W_n(\omega) = \sum_{l=0}^{L-1} w_{n,l} e^{-j\omega l}$ is the Discrete-Time Fourier Transform (DTFT) of \mathbf{w}_n , and $g_n(\omega, \phi, \theta)$ represents the sensor response of the n -th sensor to a plane wave with frequency ω originating from direction (ϕ, θ) . Here, ϕ and θ denote azimuth and elevation angle and are measured with respect to the positive x - and z -axis, respectively, as in [9].

In [1], we proposed the design of an HRTF-based RLSFI FSB

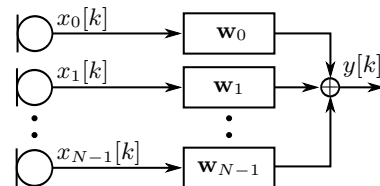


Fig. 1. Illustration of filter-and-sum beamforming [4].

where a desired beamformer response $\hat{B}(\omega, \phi, \theta)$ is approximated in the Least-Squares (LS) sense at each frequency ω . In addition, a distortionless response constraint in the desired look direction and a lower bound on the WNG are imposed on the filter coefficients. The LS approximation is performed for a discrete set of Q frequencies ω_q and M look directions (ϕ_m, θ_m) , and can be formulated in matrix notation as

$$\underset{\mathbf{w}_f(\omega_q)}{\operatorname{argmin}} \|\mathbf{G}(\omega_q) \mathbf{w}_f(\omega_q) - \hat{\mathbf{b}}\|_2^2 \quad (2)$$

subject to:

$$\frac{|\mathbf{w}_f^T(\omega_q) \mathbf{d}(\omega_q)|^2}{\mathbf{w}_f^H(\omega_q) \mathbf{w}_f(\omega_q)} \geq \gamma > 0, \quad \mathbf{w}_f^T(\omega_q) \mathbf{d}(\omega_q) = 1, \quad (3)$$

where $\mathbf{w}_f(\omega_q) = [W_0(\omega_q), \dots, W_{N-1}(\omega_q)]^T$, $[\mathbf{G}(\omega_q)]_{mn} = g_n(\omega_q, \phi_m, \theta_m)$, $\hat{\mathbf{b}} = [\hat{B}(\phi_0, \theta_0), \dots, \hat{B}(\phi_{M-1}, \theta_{M-1})]^T$ is a vector which contains the desired response for all M look directions, and $\mathbf{d}(\omega_q) = [g_0(\omega_q, \phi_{\text{ld}}, \theta_{\text{ld}}), \dots, g_{N-1}(\omega_q, \phi_{\text{ld}}, \theta_{\text{ld}})]^T$ is the steering vector corresponding to the desired look direction $(\phi_{\text{ld}}, \theta_{\text{ld}})$. Moreover, operators $\|\cdot\|_2$, $(\cdot)^T$, and $(\cdot)^H$ denote the Euclidean norm, and the transpose and conjugate transpose of vectors or matrices, respectively. Note that the same desired response is chosen for all frequencies, as can be seen from the frequency-independent entries of $\hat{\mathbf{b}}$, hence the term frequency-invariant beamformer design [4]. Equations (2) and (3) can be interpreted as follows: The LS approximation of the desired beamformer response is given by (2). The first part of (3) represents the WNG constraint with the lower bound γ on the WNG, which is a user-defined parameter [4]. The second part of (3) describes the distortionless response constraint, which ensures that the target signal originating from the desired look direction passes the beamformer undistorted. After solving the convex optimization problem in (2), (3) for each frequency ω_q separately, the time-domain FIR filters \mathbf{w}_n are obtained by an FIR approximation of the resulting optimum frequency response samples $\mathbf{w}_f(\omega_q)$. To solve the optimization problem, we used CVX, a package for specifying and solving convex optimization problems [10, 11].

In order to account for the scattering effects of the humanoid robot's head on the sound field, we use HRTFs as steering vectors in the optimization problem. The HRTFs need to be measured for the microphone array on the robot's head beforehand. Hence, $g_n(\omega_q, \phi_m, \theta_m)$ in (2) and (3) are given by

$$g_n(\omega_q, \phi_m, \theta_m) = h_{mn}(\omega_q), \quad (4)$$

where $h_{mn}(\omega_q)$ denotes the HRTF from the m -th direction to the n -th microphone at the q -th frequency.

In the following, we present a design example of the HRTF-based RLSFI beamformer design, which was also employed in our previous work [1]. We will then make use of this design example to motivate the necessity of a two-dimensional beamformer design. The beamformer design was carried out for the 12-microphone head array shown in Fig. 7(a). For the design, we used the one-dimensional desired response illustrated in Fig. 2 for each frequency,

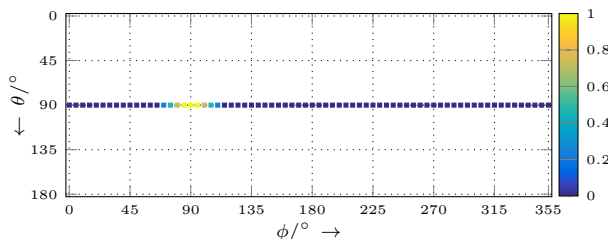


Fig. 2. One-dimensional desired response $\hat{\mathbf{b}}$ for HRTF-based RLSFI beamformer illustrated in Figs. 3 and 4.

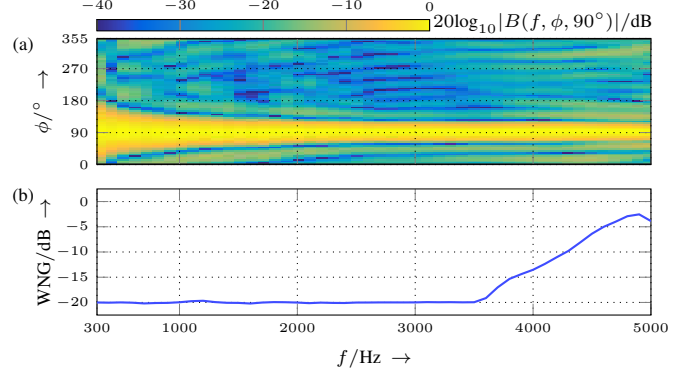


Fig. 3. Illustration of (a) beampattern and (b) WNG of the HRTF-based beamformer, designed with the one-dimensional desired beamformer response in Fig. 2 with $(\phi_{\text{ld}}, \theta_{\text{ld}}) = (90^\circ, 90^\circ)$ and $10\log_{10}\gamma = -20$ dB.

where each square represents one direction (ϕ_m, θ_m) for which the desired beamformer response is specified. The actual value of $\hat{B}(\phi_m, \theta_m)$ is coded by each square's color. Hence, the desired response in Fig. 2 has been defined for azimuth angles between 0° and 355° in steps of 5° and for a fixed elevation angle of $\theta_m = 90^\circ, \forall m$. It is equal to one for the target look direction $(\phi_{\text{ld}}, \theta_{\text{ld}}) = (90^\circ, 90^\circ)$ and decreases to zero at both sides, with a 3-dB beamwidth of 20° . Furthermore, we used a lower bound of $10\log_{10}\gamma = -20$ dB and an FIR filter length of $L = 1024$ for the beamformer design. The HRTFs $h_{mn}(\omega_q)$ which were used for the beamformer design were measured beforehand (see Section 3 for more details on the measurement process).

The resulting beampattern (normalized such that the maximum is equal to 0 dB) and WNG is illustrated in Figs. 3(a) and 3(b), respectively, for a frequency range of $300 \text{ Hz} \leq f \leq 5000 \text{ Hz}$ (chosen with the application to speech signal capture in mind). Note that the beampattern was computed with HRTFs modeling the acoustic system (4). Thus, it effectively shows the transfer function between source position and beamformer output. As can be seen, the beampattern exhibits a very narrow main beam for $f \geq 1500 \text{ Hz}$, whereas below that frequency, the main beam widens. It can also be seen that the design fulfills the WNG constraint with minor deviations which are due to the FIR approximation of the optimum filter coefficients with finite length.

2.2. Extension to two dimensions

Employing a three-dimensional microphone array as the one in Fig. 7(a) offers the capability to distinguish between sound waves coming from all directions around the robot's head. Therefore, in order to assess the quality of a beamformer design for a three-dimensional array, one has to consider the beamformer response for the three-dimensional sound field. Fig. 2 already indicates that when using the illustrated one-dimensional desired response, the beamformer's behavior is only controlled in a plane corresponding to a fixed elevation angle (here, $\theta = 90^\circ$), and not for the entire surrounding sound field. In Fig. 4, we illustrate the complete beampattern of the one-dimensional beamformer design, now depending on both azimuth and elevation angle, i.e., for all steering angles on a sphere around the array. Since the complete beampattern at every single frequency is two-dimensional, we only show it for two different frequencies $f \in \{1000 \text{ Hz}, 3000 \text{ Hz}\}$. The red rectangular areas denote the elevation angle for which the beamformer design was controlled by specifying the one-dimensional desired response

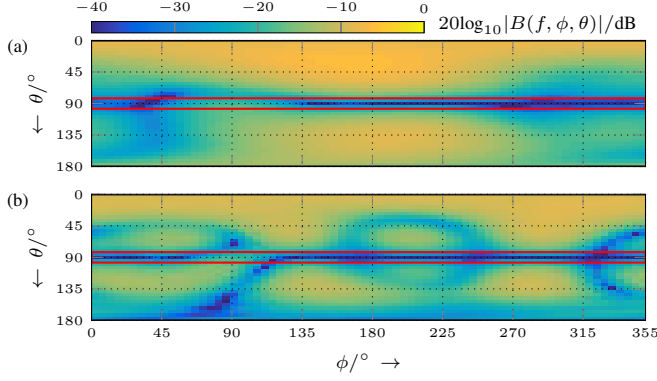


Fig. 4. Illustration of complete beampatterns of the one-dimensional beamformer design with desired beamformer response in Fig. 2 at frequencies (a) $f = 1000$ Hz and (b) $f = 3000$ Hz, with $(\phi_{ld}, \theta_{ld}) = (90^\circ, 90^\circ)$ and $10\log_{10}\gamma = -20$ dB.

in Fig. 2. Note that the parts of the beampatterns in Figs. 4(a) and 4(b) within these areas, correspond to vertical slices through the beampattern in Fig. 3(a) at the respective frequencies. The different scaling is due to the fact that the beampattern in Fig. 3(a) was normalized to the maximum value of the beampattern in the design plane corresponding to $\theta = 90^\circ$, whereas the beampatterns in Fig. 4 were normalized to the global maximum value of all beampatterns for all angles and frequencies. It can be seen that within the red areas the beamformer design is fulfilled, since the beampatterns exhibit a local maximum in the target look direction $(\phi_{ld}, \theta_{ld}) = (90^\circ, 90^\circ)$. However, large maxima in angular regions where the beamformer design was not controlled, i.e., where no desired response was defined, can be observed. Consequently, this beamformer cannot be expected to work well in a practical scenario, and it is thus necessary to design the beamformer for the entire surrounding sound field.

For this purpose, we propose to define a two-dimensional desired response along azimuth and elevation angle such that the beamformer's behavior can be controlled for the entire three-dimensional sound field surrounding the robot. Due to the two-dimensional desired response, we refer to this design as the two-dimensional beamformer design. An exemplary desired response is illustrated in Fig. 5. Here, we defined the desired response for the entire angular region in steps of five degrees in both azimuth and elevation direction, except for $\theta \in \{0^\circ, 180^\circ\}$, where we defined the desired response for $\phi = 90^\circ$ only. When using a two-dimensional desired beamformer response, the first dimensions of $\mathbf{G}(\omega_q)$ and $\hat{\mathbf{b}}$ in (2) and (3) increase due to the larger number of look directions M . However, the optimization problem is still convex. Hence, we follow the same approach to solve it as before.

The resulting beampatterns and the WNG are illustrated in Fig. 6. The beampatterns now exhibit a distinct global maximum in the target look direction. Moreover, the WNG constraint is still fulfilled. Thus, the extended two-dimensional beamformer design

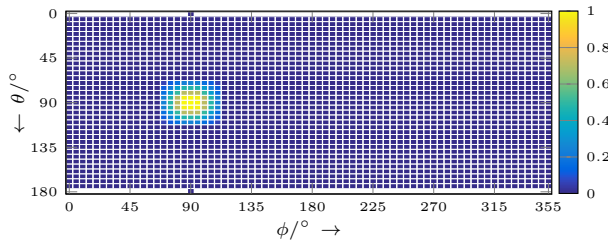


Fig. 5. Two-dimensional desired response $\hat{\mathbf{b}}$ for HRTF-based RLSFI beamformer illustrated in Fig. 6.

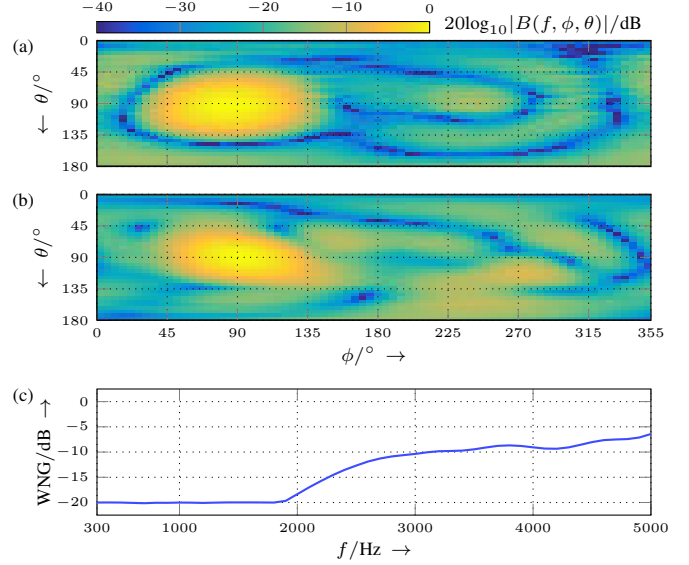


Fig. 6. Illustration of complete beampatterns of the two-dimensional beamformer design with desired beamformer response in Fig. 5 at frequencies (a) $f = 1000$ Hz and (b) $f = 3000$ Hz, with $(\phi_{ld}, \theta_{ld}) = (90^\circ, 90^\circ)$ and $10\log_{10}\gamma = -20$ dB. The corresponding WNG is shown in sub-figure (c).

still yields a feasible solution and should now be applicable to a practical robot audition scenario.

3. EXPERIMENTS

In the following, we evaluate the one- and two-dimensional beamformer designs in a robot audition scenario, and compare their respective signal enhancement performances.

As for the design examples, we used a lower bound on the WNG of $10\log_{10}\gamma = -20$ dB and FIR filters of length $L = 1024$ samples at a sampling rate of $f_s = 16$ kHz. The beamformers were designed for the 12-microphone array illustrated in Fig. 7(a), which was developed during the Embodied Audition for RobotS (EARS) project [12]. The microphone positions were chosen such that spatial aliasing for low Spherical Harmonics (SH) orders is significantly reduced [13]. In combination with mechanical constraints, this led to the seemingly random distribution of microphones. The HRTFs for the beamformer design were measured in a low-reverberation chamber ($T_{60} \approx 50$ ms) for 2522 loudspeaker positions distributed on a sphere with a radius of 1.1m around the robot's head (with discrete steps of five degrees in azimuth and elevation direction), using maximum-length sequences, see, e.g., [14]. Due to mechanical constraints, the HRTFs had to be measured without the robot's body,

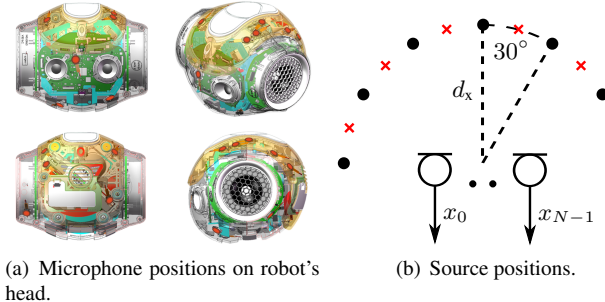


Fig. 7. Illustration of microphone positions (red circles) at the 12-microphone humanoid robot's head, and of the evaluated scenario.

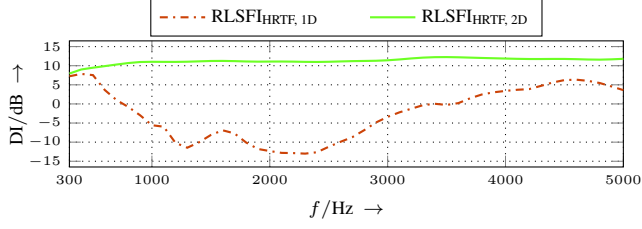


Fig. 8. Directivity index of the one- (red dash-dotted curve) and two-dimensional (green curve) HRTF-based beamformers with $(\phi_{\text{ld}}, \theta_{\text{ld}}) = (90^\circ, 90^\circ)$ and $10\log_{10}\gamma = -20\text{dB}$.

i.e., the robot's head was mounted on a microphone stand. Analysis of the influence of the robot's body on the measured HRTFs is an aspect of future work.

At first, we evaluate the two beamformer designs with respect to their Directivity Index (DI), see, e.g., (2.19) in [6], which is illustrated in Fig. 8. As can be seen, the one-dimensional beamformer design only yields a very limited DI across the entire frequency range (red dash-dotted curve), with a maximum of approximately 8 dB at 400 Hz and a minimum of -13 dB at 2300 Hz. For $400\text{ Hz} \leq f \leq 3400\text{ Hz}$ the DI is below 0 dB. Clearly, this beamformer is not suited for a practical scenario. In comparison, the two-dimensional beamformer design yields a much higher DI for all frequencies (green curve), with the DI being above 11 dB for almost all frequencies, with a maximum of 12.3 dB at 3500 Hz.

Second, we evaluate the signal enhancement performance using the frequency-weighted segmental Signal-to-Noise Ratio (fwSegSNR) according to (8) in [15]. The fwSegSNR at the beamformer's input and output was calculated using the desired signal components at the frontmost microphone and at the beamformer's output as reference signal, respectively. The two beamformers were evaluated in a two-speaker scenario, which is illustrated in Fig. 7(b), where target and interfering sources are represented by black circles and red crosses, respectively. The target source was located at positions between $\phi_{\text{ld}} = 0^\circ$ and $\phi_{\text{ld}} = 180^\circ$ in steps of 30° , at an elevation angle of $\theta_{\text{ld}} = 90^\circ$. The direction of arrival of the target source was assumed to be known, i.e., no localization algorithm was applied. For each target position, seven interfering speaker positions between $\phi_{\text{int}} = 15^\circ$ and $\phi_{\text{int}} = 165^\circ$ in steps of 30° were evaluated. During the first experiment, the interfering sources were located at an elevation angle of $\theta_{\text{int}} = 90^\circ$, whereas in the second experiment $\theta_{\text{int}} = 73^\circ$. The target source was always located at an elevation angle of 90° . Note that we chose the two different elevation angles for the interfering sources in order to demonstrate the situation, when target source and interferer are not in the same plane for which the one-dimensional beamformer was designed. The fwSegSNR was calculated for each combination of target and interfering source positions and averaged over the fwSegSNR values obtained for the different interferer positions. The resulting average target source position-specific fwSegSNR levels are illustrated in Fig. 9, where Figs. 9(a) and 9(b) depict the results for $(\theta_{\text{int}} = 90^\circ)$ and $(\theta_{\text{int}} = 73^\circ)$, respectively. The microphone signals were created by convolving clean speech signals of duration 366 s (200 concatenated utterances from the GRID corpus [16]) with head-related Room Impulse Responses (RIRs), which were measured in a lab room with a reverberation time of $T_{60} \approx 400\text{ ms}$, at a horizontal distance between robot head (at a height of 1.2 m) and loudspeaker of $d_x = 1\text{ m}$ and a vertical distance of either $d_z = 0\text{ m}$ ($\theta_{\text{int}} = 90^\circ$) or $d_z = 0.3\text{ m}$ ($\theta_{\text{int}} = 73^\circ$), respectively.

In general, one can observe that when the interfering source is located at the same elevation angle as the target source, the input fwSegSNR levels as well as most of the output fwSegSNR levels are

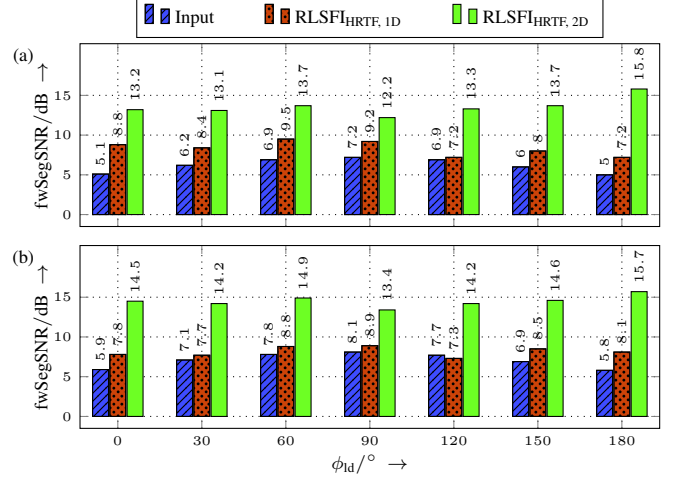


Fig. 9. Average target source position-specific fwSegSNRs in dB, obtained at the input (blue bars) and at the output of the one- (red bars) and two-dimensional (green bars) HRTF-based RLSFI beamformers. Results were obtained for $T_{60} \approx 400\text{ ms}$ with interfering sources at an elevation of (a) $\theta_{\text{int}} = 90^\circ$ and (b) $\theta_{\text{int}} = 73^\circ$.

lower compared to $\theta_{\text{int}} = 73^\circ$, i.e., when the interfering source is located above the target source. One reason for this might be that the elevation angle of 73° corresponds to a larger source-robot distance than for an elevation angle of 90° , since the RIRs for both elevation angles were measured for a fixed horizontal distance between robot and loudspeaker. When looking at the results of the one-dimensional beamformer design, we observe a slight improvement of the signal enhancement for $\theta_{\text{int}} = 90^\circ$. When the interfering sources are no longer located in the same plane as the target source, the average fwSegSNR gain for the one-dimensional design is lower for most of the evaluated target look directions, see, Fig. 9(b). This demonstrates the drawback of the one-dimensional beamformer design, being only designed for a specific elevation angle. This is not the case for the two-dimensional beamformer design, which consistently improves the fwSegSNR to a great extent for all tested look directions and elevation angles. A closer look reveals that the best average performance was obtained for $\phi_{\text{ld}} = 180^\circ$, which we think is due to the distribution of microphones on the robot's head, where more microphones are located on the left-hand side of the head than on the right-hand side.

To summarize, the results confirm the effectiveness of the two-dimensional beamformer design in a realistic robot audition scenario, in which the previous one-dimensional beamformer design does not provide acceptable signal enhancement performance.

4. CONCLUSION

In this work, we proposed a two-dimensional HRTF-based RLSFI beamformer design for robot audition. As opposed to the previously published one-dimensional HRTF-based RLSFI beamformer design, we now explicitly control the beamformer response for the entire sphere of possible DoAs around the robot's head. We evaluated both beamformer designs with respect to the DI and their corresponding signal enhancement performance, which was evaluated by means of fwSegSNR levels. The results confirmed the effectiveness of the two-dimensional beamformer design, which resulted in a higher DI and an increased and more reliable signal enhancement performance in a typical robot audition scenario, compared to the one-dimensional beamformer design.

Future work includes an investigation of the proposed two-dimensional beamformer design with respect to different desired responses as well as an extension to polynomial beamforming [17, 18] to allow for flexible beam steering in all directions.

5. REFERENCES

- [1] H. Barfuss, C. Huemmer, G. Lamani, A. Schwarz, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant beamforming," in *IEEE Workshop Appl. Signal Process. Audio Acoustics (WASPAA)*, Oct. 2015, pp. 1–5.
- [2] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Blind source separation for robot audition using fixed HRTF beamforming," *EURASIP J. Advances Signal Proc. (JASP)*, vol. 2012, pp. 58, 2012.
- [3] M. Maazaoui, Y. Grenier, and K. Abed-Meraim, "From bin-audal to multimicrophone blind source separation using fixed beamforming with HRTFs," in *Proc. IEEE Int. Conf. Systems, Signals, Image Process. (IWSSIP)*, Apr. 2012, pp. 480–483.
- [4] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 77–80.
- [5] E. Mabande, *Robust Time-Invariant Broadband Beamforming as a Convex Optimization Problem*, Ph.D. thesis, Friedrich-Alexander University Erlangen-Nürnberg, Apr. 2014.
- [6] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*, pp. 19–38. Springer Berlin Heidelberg, 2001.
- [7] W. Herboldt, *Sound capture for human/machine interfaces: Practical aspects of microphone array signal processing*, vol. 315, Springer Science & Business Media, 2005.
- [8] H. Cox, R. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process. (ASSP)*, vol. 34, no. 3, pp. 393–398, June 1986.
- [9] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, Detection, Estimation, and Modulation Theory. Wiley, 2004.
- [10] Inc. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0," <http://cvxr.com/cvx>, Aug. 2012.
- [11] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds., Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.
- [12] Seventh Framework Programme, "Embodied Audition for RobotS (EARS)," Oct. 2016, <http://robot-ears.eu/>.
- [13] V. Tourbabin and B. Rafaely, "Design of pseudo-spherical microphone array with extended frequency range for robot audition," in *Jahrestagung für Akustik (DAGA)*, Aachen, Germany, Mar. 2016, pp. 1068–1071.
- [14] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Am. (JASA)*, vol. 66, no. 2, Aug. 1979.
- [15] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process. (ASL)*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
- [16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am. (JASA)*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.
- [17] M. Kajala and M. Hamalainen, "Filter-and-sum beamformer with adjustable filter characteristics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May 2001, pp. 2917–2920.
- [18] E. Mabande and W. Kellermann, "Design of robust polynomial beamformers as a convex optimization problem," in *Proc. IEEE Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Aug. 2010, pp. 1–4.